In Proceedings, *Workshop "New Approaches to Learning for NLP" at 14th International Joint Conference on Artificial Intelligence (IJCAI'95)*, Montréal. *Forthcoming.*

A Symbolic and Surgical Acquisition of Terms through Variation*

Christian Jacquemin
Institut de Recherches en Informatique de Nantes (IRIN)
IUT de Nantes, 3 rue M^{al} Joffre, 44041 Nantes, France
Phone : +33 40 30 60 52 — Fax : +33 40 30 60 53
jacquemin@irin.iut-nantes.univ-nantes.fr

Abstract

Terminological acquisition is an important issue in learning for NLP due to the constant terminological renewal through technological changes. Terms play a key role in several NLP-activities such as machine translation, automatic indexing or text understanding. In opposition to classical once-and-for-all approaches, we propose an incremental process for terminological enrichment which operates on existing reference lists and large corpora. Candidate terms are acquired by extracting variants of reference terms through *FASTR*, a unification-based partial parser. As acquisition is performed within specific morpho-syntactic contexts (coordinations, insertions or permutations of compounds), rich conceptual links are learned together with candidate terms. A clustering of terms related through coordination yields classes of conceptually close terms while graphs resulting from insertions denote generic/specific relations. A graceful degradation of the volume of acquisition on partial initial lists confirms the robustness of the method to incomplete data.

Aims

Multi-word terms and compounds play an increasing role in language analysis for the following reasons : their interpretation is rarely transparent, they generally denote a specific class of mental or real-world objects and the words composing them are strongly related. Therefore, a correct processing of terms ensures a higher quality in several applications of Natural Language Processing (NLP). In Machine Translation, their lack of transparency makes word-for-word translation fail and calls for specific descriptions. In Information Retrieval, their high informational content makes them good descriptors (Lewis & Croft 1990). In parsing, the selectional restrictions found between head words and their arguments within a term give important clues for structural noun phrase disambiguation (Resnik 1993).

As terms mirror the concepts of the domain to which they belong, a constant knowledge evolution leads to a constant term renewal. Thus terminological acquisition is a necessary companion to NLP, specifically when dealing with technical texts.

Tools for terminological acquisition, whether statistical, such as (Church & Hanks 1989), or symbolic, such as (Bourigault 1993), acquire terms from large corpora through a once-and-for-all process without consideration for any prior terminological knowledge. This lack of incrementality in acquisition has the following drawbacks :

- Acquired terms must be merged with the initial ones with consideration of eventual variants.
- Acquired terms are neither conceptually nor linguistically related to the original ones.
- The set of original terms is ignored although it could be a useful source of knowledge for acquisition.

It is possible to conceive a finer approach to term acquisition by considering the local variants of terms within corpora. As term variants generally involve more than one term, their extraction can fruitfully exploit existing lists of terms in a process of non massive incremental acquisition.

*We would like to thank the French scientific documentation center *INIST/CNRS* for providing us with data. All the experiments reported in this paper have been performed on [Pascal] a list of 71,623 multi-domain terms and [Medic] a 1.56-million word medical corpus composed of abstracts of scientific papers owned by *INIST/CNRS*. Many thanks to Jean Royauté of *INIST/CNRS* for his helpful and friendly collaboration. This work has also benefited from rich discussions in the research group *Terminologie et Intelligence Artificielle* of the *PRC Intelligence Artificielle*.

For example, if *viral hepatitis* is a known term, *viral and autoimmune hepatitis* is a variant of this term (a coordination) which displays *autoimmune hepatitis* as a candidate term. Moreover, this coordination indicates a strong closeness between the interpretation of both terms which can be associated to a link within a thesaurus. Henceforth, potential terms acquired through acquisition techniques will be called *candidate terms*. The decision whether to include a candidate term into a terminology is outside the scope of our work.

Acquiring with a Concern for Prior Knowledge

Tools for acquiring terms generally operate on large corpora using various techniques to detect term occurrences. There are mainly two families of tools for term acquisition : statistical measures and NLP symbolic techniques.

The first family which comprises most of the tools is composed of statistical analyzers which have little or no linguistic knowledge. These applications take advantage of the specific statistical behavior of words composing terms : words which are lexically related tend to be found simultaneously more frequently than they would be just by chance. Pure statistical methods such as (Church & Hanks 1989) are rare. Generally some linguistic knowledge is associated to the statistical measures through a prior (Daille 1994) or a posterior (Smadja 1993) filtering of correct syntactic patterns. The assumption implicitly stated by statistical works, and which is backed up by our study, is that it is more likely to find a term in the neighborhood of another one than anywhere else in a text. More specifically, we assume that the best way of combining two terms syntactically and semantically is to build a specific structure that we call a *variant* which is either a term or a restricted noun phrase and which is observed within a small span of words.

The second approach to term acquisition consists of knowledge-based methods which rely on local grammars of noun phrases and compounds (Bourigault 1993). Word sequences accepted by these grammars are extracted through a more or less shallow parse of corpora and are good candidate terms.

The counterpart of both statistical and knowledge-based acquisitions is to provide the user with large lists of candidates which have to be manually filtered out. For example, *LEXTER* (Bourigault 1993) extracts 20,000 occurrences from a 200,000-word corpus which represent 10,000 candidate terms. It is due to a lack of initial terminological knowledge along with a lack of consideration for terminological variation that such methods propose too large sets of terms. In order to reduce the volume of acquisition and also to propose candidates which are more likely to be terms, this paper presents a method based on an initial list of terms called *reference terms*. The acquisition procedure starts from this supposed comprehensive set of reference terms. It decomposes variations of these terms found in corpora and is then able to detect candidate terms.

Updating Rather Than Acquiring

Is it realistic to suppose that lists of terms exist for technical domains ? The ever-growing mass of electronic documents calls for tools for accessing these data which have to make extensive use of term lists as sources of indexes. For this purpose, and for other activities related to textual databases, more and more thesauri exist. Some of them, such as the *Unified Medical Language System* meta-thesaurus, carry conceptual and/or linguistic information about the terms they contain. In our experiment we have used the [Pascal] terminological list composed of 71,623 multi-domain terms without conceptual links, provided by the documentation center *INIST/CNRS*.

Because of the availability of large term lists, it is natural to lay a greater stress on the updating of such data than on their acquisition from scratch. Therefore, our approach to acquisition focuses on how to improve a list of terms through the observation of a corpus. Our approach also differs from previous experiments on term acquisition because it yields conceptual links between candidate and reference terms. It can be used to check or to enhance the conceptual knowledge of thesauri in a way complementary to automatic semantic clustering of terms through an observation of their syntactic contexts (Grefenstette 1994).

A Micro-syntax for an Accurate Extraction

The first step in our approach to terminological acquisition is the extraction of term variants from a large corpus. The tool used is *FASTR*, a unification-based partial parser. *FASTR* recycles lists of reference terms by transforming them into grammar rules. Then, it dynamically builds term variant rules from these term rules. The parser is described in (Jacquemin 1994) and, here,

it will just be sketched out, by focusing on the features that are relevant for terminological acquisition. More specifically, we will omit the aspects of the parser concerning its optimization and the feature structures associated with rules and meta-rules.

In such a simplified framework, each reference term corresponds to a *PATR-II*-like rule (Shieber 1986) comprising a context free skeleton and lexical items. For example, rule (1) denotes the term *serum albumin* with a $\langle Noun \rangle \langle Noun \rangle$ structure :

$$\begin{array}{ll} \text{Rule} & N_1 \rightarrow N_2 \; N_3 \;\; : \\ & < N_2 \; lemma > \doteq \text{`}serum\text{'} \\ & < N_3 \; lemma > \doteq \text{`}albumin\text{'}. \end{array} \tag{1}$$

At a higher level, a set of meta-rules operates on the term rules and produces new rules describing potential variations. Each meta-rule is dedicated to a specific term structure and to a specific type of variation. For the sake of clarity, meta-rules are divided into two sets – meta-rules for two-word terms and meta-rules for three-word terms – and each set is subdivided into three subsets – meta-rules for coordination, insertion and permutation. Meta-rules for terms of four words or more are ignored because they produce very few variants (approximately 1 % of the variants). Meta-rule (2) applies to rule (1) and yields a new rule (3) :

$$\begin{array}{ll} \text{Metarule} & Coor(X_1 \rightarrow X_2 \; X_3) \\ & \equiv \;\; X_1 \rightarrow X_2 \; C_4 \; X_5 \; X_3 \;\; : . \end{array} \tag{2}$$

$$\begin{array}{ll} \text{Rule} & N_1 \rightarrow N_2 \; C_4 \; X_5 \; N_3 \;\; : \\ & < N_2 \; lemma > \doteq \text{`}serum\text{'} \\ & < N_3 \; lemma > \doteq \text{`}albumin\text{'}. \end{array} \tag{3}$$

This transformed rule accepts any sequence *serum* C_4 X_5 *albumin* as a variant of *serum albumin* where C_4 is any coordinating conjunction and X_5 any single word. For example, it correctly recognizes *serum and egg albumin* as a variant of *serum albumin*. The second column of Table 1 presents some other meta-rules for two-word terms together with examples of pairs composed of a term and one of its variants. Currently, the meta-grammar of *FASTR* for English includes 73 meta-rules for 2- and 3-word terms : 25 coordination meta-rules, 17 insertion meta-rules and 31 permutation meta-rules (plus 66 meta-rules for 4-word terms which are not used for acquisition).

When term variants are described through meta-rules as in *FASTR*, it is very simple to devise a process for term acquisition : each paradigmatic meta-rule (or skeleton of a filtering meta-rule) is linked to a pattern extractor, yielding a candidate term. As no further analysis of the variants is required, such an acquisition is extremely fast. The acquisition of terms by extracting patterns from variants is processed as follows for the different categories of variants :

- *Coordination.* The candidate term is the term coordinated with the original one.
- *Insertion.* The candidate term is the term which has replaced the head of the original term through substitution.
- *Permutation.* In a permutation of a 2-word term, the argument of the original term is shifted from the left of the head to its right and is transformed into a prepositional phrase. The candidate term is the noun phrase inside this prepositional phrase. This definition is extended to terms of 3 words or more where one of the arguments is permuted.

The third column of Table 1 exemplifies patterns of acquisition for each of the three categories of term variants.

This method for term acquisition does not systematically succeed for each encountered term variant. Some correct variants involve only one term instead of two or more and cannot produce new candidates. For example, *cells and their subpopulations* is a coordination variant of *cell subpopulation* which is unproductive compared with the variant exemplified for coordination in Table 1. Moreover, terms acquired through a variation may already be reference terms (see the non-underlined candidates in Tables 2, 3 and 4). For the reference list to be sufficiently comprehensive, it is expected and even desirable that some of the acquired terms are already known. Moreover, “acquisitions” of known terms are not useless because they reveal conceptual links between these terms.

Acquiring Conceptual Classes

Tables 2, 3 and 4 exemplify some terms acquired through the three main kinds of variations observed for English : coordinations, insertions and permutations. The terms acquired through permutations are not conceptually related to the original ones due to the syntagmatic nature of this transformation. On the contrary, coordination and insertion variations relate semantically close terms. We examine in turn the decomposition of these two kinds of variations in the aim of acquiring conceptual links.

	Meta-rule and associated variant	Acquisition
Coordination (25 meta-rules)	$X_2\ X_3 \mapsto X_2\ X_4\ C_5\ X_3$ *surgical closure* $\mapsto$ *surgical exploration and closure*	$X_2\ X_4$ *surgical exploration*
Insertion (17 meta-rules)	$X_2\ X_3 \mapsto X_2\ X_4\ X_3$ *medullary carcinoma* $\mapsto$ *medullary thyroid carcinoma*	$X_4\ X_3$ *thyroid carcinoma*
Permutation (31 meta-rules)	$X_2\ X_3 \mapsto X_3\ P_4\ X_5\ X_2$ *control center* $\mapsto$ *center for disease control*	$X_5\ X_2$ *disease control*

Table 1: Acquisition through pattern extraction from variants. (Examples are from [Medic].)

Candidate term	Reference Term
abdominal aorta	*Thoracic aorta*
acidic lipid	*Neutral lipid*
active phase	*Latent phase*
adrenal gland	*Thyroid gland*
affective disorder	*Cognitive disorder*
aged animal	*Young animal*
agonist bromocriptine	*Agonist antagonist*
air conduction	*Bone conduction*
amniotic fluid estimation	*Ratio estimation*
aortic arch	*Aortic coarctation*
aortic valve	*Mitral valve*
arterial acid base	*Arterial blood*

Table 2: Examples of term acquisition through coordination from [Medic]. Terms which do not belong to the reference list are underlined.

Candidate term	Reference Term
abdominal spear injury	*Penetrating injury*
ablating tool	*Cutting tool*
absorbed dose	*Radiation dose*
access pressure	*Blood pressure*
accessory nerve	*Spinal nerve*
acetylcholine receptor	*Muscarinic receptor*
acetylcholine receptor	*Nicotinic receptor*
acid analysis	*Organic analysis*
acid base disorder	*Metabolic disorder*
action potential	*Evoked potential*
action potential	*Membrane potential*
activity curve	*Time curve*

Table 3: Examples of term acquisition through insertion from [Medic]. Terms which do not belong to the reference list are underlined.

Coordination

Two terms are coordinated only if they share the same semantic scheme. For example, the variant *surgical exploration and closure* (see the first example of Table 1) indicates that the two terms *surgical exploration* and *surgical closure* are semantically close. They both denote a surgical act. This fact is interesting because some of the terms with a *surgical* ⟨*Noun*⟩ structure such as *surgical shock* do not belong to the same conceptual class and could not be coordinated with any of the *surgical* ⟨*Noun*⟩ terms from this class : **a surgical shock and closure* is incorrect. Thus, when heads are coordinated (approximately 15 % of the coordinations) the head nouns of the terms must belong to the same semantic class (with respect to their entry selected by their argument). On the other hand, when arguments are coordinated, they must select the same entry of the head noun. For example, *dorsal spine* and *cervical spine* can be coordinated as both being a part of the *(nervous) spine* but neither of them can be coordinated with a *hedgehog* or a *fish spine*. Such coordinations are useful indicators for the disambiguation of a head word by its arguments :

- For its classification with other related words through head coordination.

- For the definition of its subsenses depending on its arguments through argument coordination.

This kind of fine-grained selectional restriction has to be completed with more general information on argument structure through long distance dependencies. Such restrictions can be acquired from statistical measures on the results of a shallow syntactic analysis and semantic tags, whether manually assigned (Basili, Pazienza, & Velardi 1993) or deduced from a thesaurus (Resnik 1993). These studies provide more general and systematic restrictions than our approach and are applied to disambiguation or parsing tasks. Our acquisition

Candidate term	Reference Term
[of] accessory cell	*Cell proliferation*
[in] acetabular growth	*Growth factor*
[of] activated b cell	*Cell differentiation*
[of] acute phase protein	*Protein synthesis*
[of] adipose tissue	*Tissue extract*
[in] adult cell	*Cell function*
[in] agarose gel	*Gel electrophoresis*
[of] airway control	*Control method*
[of] anaphylatoxin level	*Level measurement*
[of] aneuploid tumor cell	*Cell population*
[of] animal tolerance	*Tolerance limit*
[of] arterial pressure	*Pressure control*

Table 4: Examples of term acquisition through permutation from [Medic]. Terms which do not belong to the reference list are underlined.

is restricted to local selection but takes advantage of the pre-existing knowledge embodied in lists of reference terms.

The acquisition from variants, illustrated for one step in Table 1, is repeated on candidate terms as long as new candidates are discovered. Then classes of compatible sense restrictions are built from terms related through constructions of coordination according to the following rule :

> Two terms t and t' are placed in the same class if and only if there exists a chain of coordination variants from t to t' : a set of n terms $t_1 = t,\ t_2, \ldots,\ t_{n-1},\ t_n = t'$ such that for each pair $(t_i, t_{i+1})_{i\in\{1,2,\ldots,n-1\}}$ either t_i is acquired from a coordination variant of t_{i+1} or t_{i+1} is acquired from a coordination variant of t_i.

Figure 1 is a planar representation of the graph constructed from one of the classes observed in the [Medic] corpus. Each arrow from a term t to a term t' indicates that t' has been acquired from a coordination variant of t.

Leaving apart the only head coordination in the figure that holds between *cirrhotic control* and *cirrhotic patient*, all the terms have a ⟨*Modifier*⟩ *control* structure [1] and can be coordinated through a head coordination. Conceptually, the terms of Figure 1 are related to a common hypernym whose linguistic utterance is *medical control*.

Moreover, the spatial organization of the graph outlines the central role played by *normal control* and *disease control*. These two terms are the most generic ones. Their root position in this acyclic graph (except for the two symmetric links) mirrors the linguistic fact that an argument coordination between two terms tends to place first the most generic argument and then the most specific one. Thus, although placed at a similar conceptual level in the taxonomy, these terms are ordered from the most generic to the most specific along the coordination links. This two-level observation reveals that linguistic clues, when precisely observed, are good indications of the conceptual organization.

Insertion

The meta-rules accounting for insertions insert one or more words inside a term string. The following meta-rule (4) denotes an insertion of one word inside a two-word term :

$$\begin{array}{ll} \text{Metarule} & Ins(X_1 \rightarrow X_2\ X_3) \\ & \equiv \quad X_1 \rightarrow X_2\ X_4\ X_3\ \ :. \end{array} \tag{4}$$

The resulting structure is ambiguous depending on whether the leftmost word of the term is still an argument of the head noun in the variation (e.g. *[inflammatory [bowel disease]]*) or an argument of the inserted word (e.g. *[[sunflower seed] oil]*). The second structure is quite rare and does not correspond to a genuine variant of the original term because it has a different argument structure. However, most of these possibly incorrect variants are correct. It happens every time when the reference term (here *sunflower oil*) corresponds to an elided denomination of the variant which is in fact the reference term. In this case, the non-ambiguity of the elided form relies on pragmatic knowledge, because everyone knows that the *seed* is the part of the *sunflower* used to make *sunflower oil*.

Whatever the structure of the variant, either $((X_3\ X_2)\ X_1)$ or $(X_3\ (X_2\ X_1))$, the extraction of the sequence $X_2\ X_1$ as candidate term (see Tables 1 and 3) yields a correct term. When extracted from the latter structure, the candidate term is more specific than the original one because modifiers in the noun phrase tend to be ordered from the most generic to the most specific.

As stated for coordination, iteration of acquisition on candidates terms yields conceptual classes. However, the construction of the graph linking terms acquired through insertion is not as straight-

[1] *Matched control* is a partial term with a missing noun argument which is not ruled out by our acquisition process. With a proper acquisition, this term would not appear as a candidate and the links issuing from this term would issue from one of the correct terms ⟨*Noun*⟩ *matched control*.

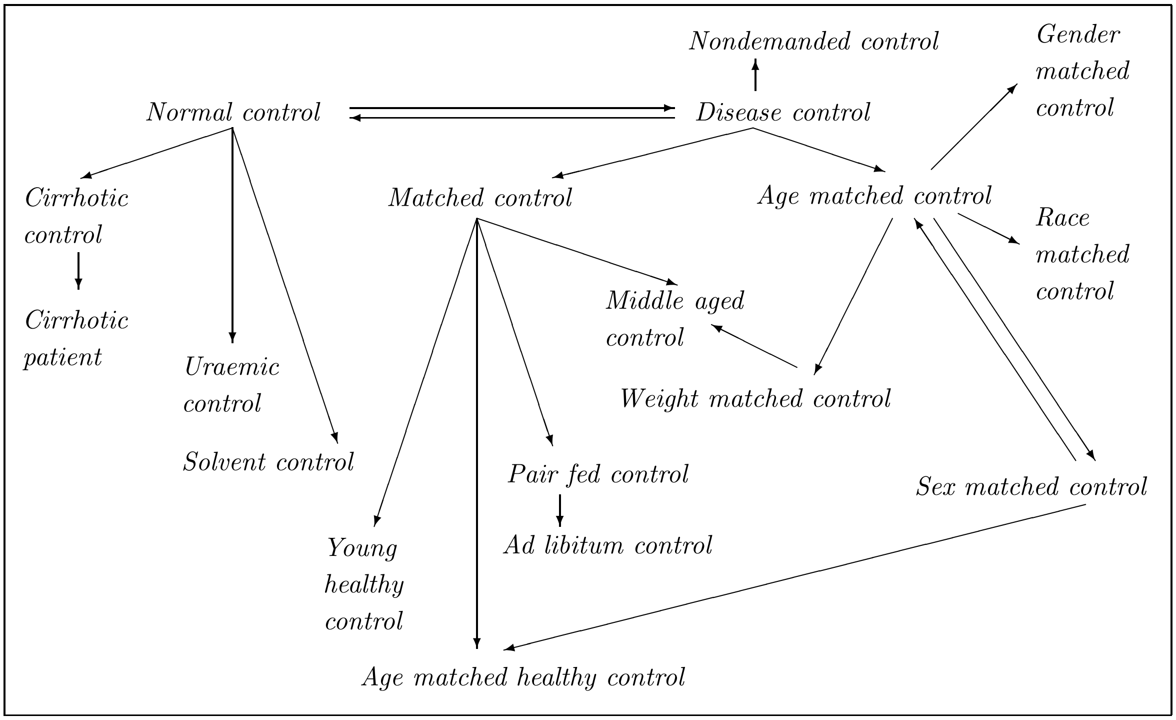

Figure 1: Network of coordination links from [Medic].

forward as it is for coordination. The reason is that one must first conflate conceptually close terms that are likely to be coordinated before constructing the hierarchy resulting from insertion variants. Figure 2 has been constructed by grouping together *malignant tumor/benign tumor*, *metastatic tumor/primary tumor* and *human tumor/experimental tumor* which have been observed in coordinated constructions. A further grouping of *rat tumor* with *human tumor* was necessary but was not indicated by a coordination in our corpus. Similarly, a general category of ⟨*Part of body*⟩ *tumors* has been created although only some coordinations were observed among the possible ones : *mammary/skin, mammary/pancreas, cutaneous/corneal, liver/lung, bone/soft tissue...*

Due to the parallel between insertion constructions and generic/specific links, there is a good similarity between the observed graph and the taxonomy of this part of the terminology. An exception to this rule is the link from ⟨*Part of body*⟩ *tumor* to *malignant tumor* coming from the variant *ovarian malignant tumor*. It is indeed an exceptional link : there are fifteen different links from *malignant tumor* to more specific terms but only one link from a more specific term (*ovarian tumor*) to *malignant tumor*.

Incrementality and Robustness

As introduced for the construction of conceptual classes, the acquisition method is repeated incrementally. Candidates are acquired from candidates of the preceding step until no new term is discovered :

> A term is a candidate if and only if there exists a chain of couples $(t_i, t_{i+1})_{i \in \{1,2,\ldots,n-1\}}$ where t_{i+1} is acquired from a variant of t_i and where t_1 is a reference term. That is to say that the set of candidates is the closure of the set of reference terms through the relation of acquisition.

Due to the finite corpus, due to the finite length of terms and due to the non circularity of the definition, the incremental acquisition reaches a fixed point after a finite number of iterations. It takes fifteen cycles to complete an acquisition of 5,080 terms when starting from the 71,623 terms of the [Pascal] list [2].

Table 5 shows five sequences of acquisition obtained from term variants in [Medic] starting from

[2] Among these 71,623 terms, only 12,717 are found in the [Medic] corpus under their basic form or one of its correct variants.

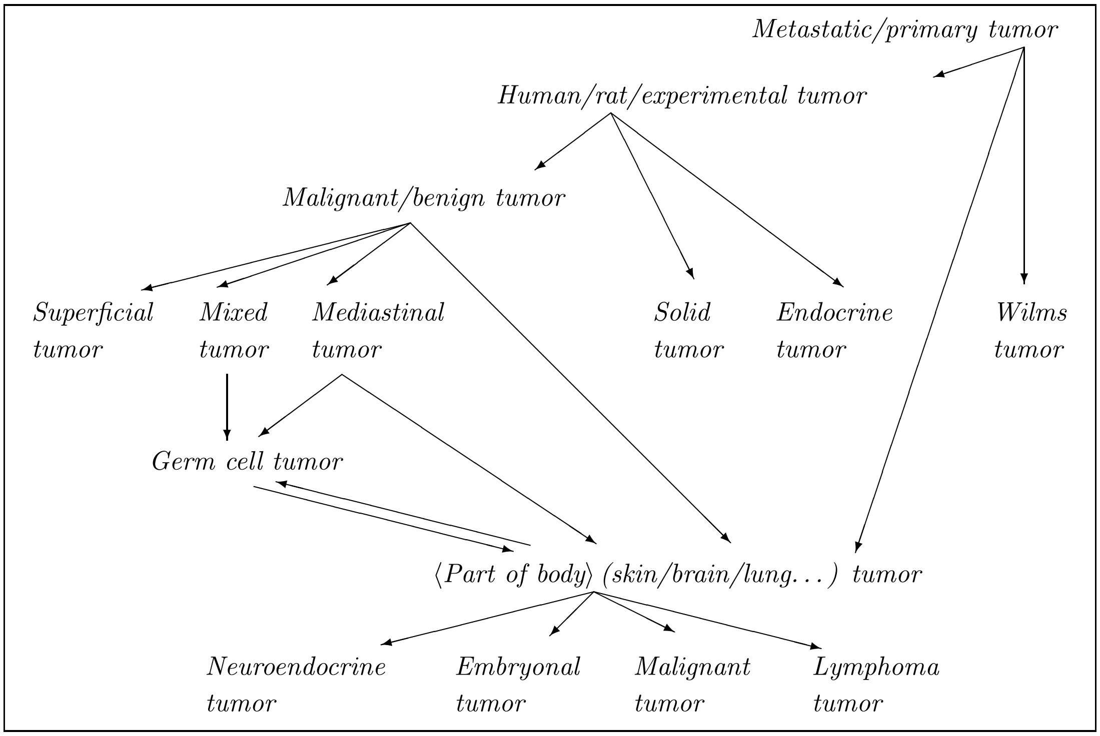

Figure 2: Network of insertion links from [Medic].

a reference term in [Pascal]. For example, the first sequence indicates the acquisition of *tumour tissue* from *tissue extract* through a permutation variant (*extract of tumour tissue*) followed by the acquisition of *normal tissue* from a coordination (*tumour or normal tissue*), and so on. This sequence mixes the three kinds of variations while the last three are restricted to insertions and/or coordinations. When not using permutation, the acquisition process yields smaller sets of terms : it produces 2,998 terms in fourteen steps through coordinations and insertions, 2,193 terms in seven steps through insertions and 357 terms in six steps through coordinations. The sets obtained without the use of permutation are "better" candidates because they are produced by transformations which yield compounds. Permutations, which transform compounds into syntactic noun phrases, tend to produce candidates of a lower quality.

As our method is based on the observation of rare occurrences, the number of acquired terms depends on the set of reference terms. As indicated in (Enguehard 1994), such a correlation does not exist in her statistical approach to term acquisition because she observes larger sets of (co-)occurrences. Figure 3 exemplifies acquisition curves for different values of the volume of reference terms. It shows that the size of the acquisition gradually degrades when the size of the bootstrap decreases : 5,080 terms are acquired when starting from the total list of 12,717 terms, 3,833 terms are still acquired from a bootstrap of 6,000 terms and 2,329 terms from a bootstrap of 1,000 terms. Thus, with only a twelfth of the initial bootstrap, almost half the terms are still acquired. Although a serious degradation of the results is observed under this lower limit, these values suggest that acquisition depends more on the size of the corpus than on the initial terminology. As a partial initial list of terms is easily compensated by a larger corpus, the completeness of the reference list is not a crucial issue for the quality of the acquisition in our framework.

Conclusion and Future Work

This study has proposed a novel approach to terminological acquisition that differs from the two main trends in this domain : morpho-syntactic filtering or statistical extraction. The main feature of our approach is accounting for existing lists of terms by observing their variants and yielding conceptual links as well as candidate terms. As long

Var	Acquired terms	
P	*Tissue extract*	*tumour tissue*
C-I	*normal tissue*	*rat tissue*
P-I	*sprague dawley rat*	*female rat*
I-I	*f344 rat*	*strain rat*
P-I	*milan strain*	*normotens. strain*
I	*control strain*	
I	*Blood cell*	*leukemic cell*
C-C	*normal cell*	*cf cell*
I-I	*pancreatic cell*	*beta cell*
C-I	*alpha cell*	*activated NK cell*
I	*Cell line*	*tumor line*
I-I	*derived cell line*	*t cell line*
I-I	*leukemia cell line*	*u937 cell line*
I	*histiocytic cell line*	
C	*Experimental study*	*clinical study*
C-C	*echocardiogr. study*	*doppler study*
C	*angiography study*	
C	*Pigment. disorder*	*nail disorder*
C-C	*nail change*	*palmar change*

Table 5: Examples of sequences of acquisition.

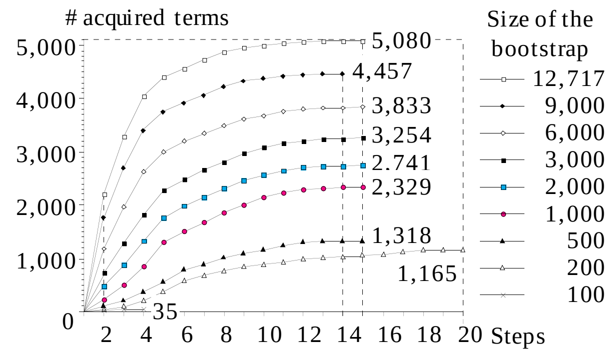

Figure 3: Acquisition volumes for different sizes of bootstrap on [Medic] corpus.

as they are accessible through morpho-syntactic dependencies in a corpus, these links can be used to automatically construct parts of the taxonomy representing the knowledge in this domain. Among the applications of this method are lexical acquisition, thesaurus discovery and technological survey. More generally, terminological enrichment is necessary for NLP activities dealing with technical sublanguages because their efficiency and their quality depend on the completeness of their lexicons of terms and compounds.

References

[Basili, Pazienza, & Velardi 1993] Basili, R.; Pazienza, M. T.; and Velardi, P. 1993. Acquisition of selectional patterns in sublanguages. *Machine Translation* 8:175–201.

[Bourigault 1993] Bourigault, D. 1993. An endogeneous corpus-based method for structural noun phrase disambiguation. In *Proceedings, 6th European Chapter of the Association for Computational Linguistics (EACL'93)*, 81–86.

[Church & Hanks 1989] Church, K. W., and Hanks, P. 1989. Word association norms, mutual information and lexicography. In *Proceedings, 27th Annual Meeting of the Association for Computational Linguistics (ACL'89)*, 76–83.

[Daille 1994] Daille, B. 1994. Study and implementation of combined techniques for automatic extraction of terminology. In *Proceedings, The Balancing Act : Combining Symbolic and Statistical Approaches to Language, Workshop at the 32nd Annual Meeting of the Association for Computational Linguistics.*

[Enguehard 1994] Enguehard, C. 1994. Acquisition of terminology from colloquial texts. In *Proceedings, Computational Linguistics for Speech and Handwriting Recognition (CLSHR).*

[Grefenstette 1994] Grefenstette, G. 1994. *Explorations in Automatic Thesaurus Discovery.* Dordrecht, The Netherlands: Kluwer Academic Publisher.

[Jacquemin 1994] Jacquemin, C. 1994. Recycling terms into a partial parser. In *Proceedings, 4th Conference on Applied Natural Language Processing (ANLP'94)*, 113–118.

[Lewis & Croft 1990] Lewis, D. D., and Croft, W. B. 1990. Term clustering of syntactic phrasess. In *Proceedings, 13th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'90)*, 385–404.

[Resnik 1993] Resnik, P. 1993. *Selection and Information : A Class-Based Approach to Lexical Relationships.* Ph.D. thesis, University of Pennsylvania, Institute for Research in Cognitive Science.

[Shieber 1986] Shieber, S. N. 1986. *An Introduction to Unification-Based Approaches to Grammar.* CSLI Lecture Notes *4*. Stanford, CA: CSLI.

[Smadja 1993] Smadja, F. 1993. Xtract : An overview. *Computer and the Humanities* 26:399–413.